

HORST LÖFGREN

THE MEASUREMENT OF
LANGUAGE PROFICIENCY

CWK GLEERUP/LUND

1972

The measurement of language proficiency

AKADEMISK AVHANDLING

som med vederbörligt tillstånd för vinnande av filosofie doktorsexamen
vid Lunds Universitets samhällsvetenskapliga fakultet
kommer att till offentlig granskning framställas i sal 122
å Lärarhögskolan i Malmö lördagen den 20 maj 1972 kl. 10.00

AV

HORST LÖFGREN

FIL. LIC., HB

HORST LÖFGREN

The measurement of language proficiency

STUDIA PSYCHOLOGICA ET PAEDAGOGICA
SERIES ALTERA XVII

HORST LÖFGREN

The measurement of language proficiency

An analysis of the structure of language proficiency,
and application of the research results in a comparison of
the effects on learning of different study materials in
the German language. A summary

LUND / CWK GLEERUP



Contents

Acknowledgements 6

Introduction 7

1. *Measuring proficiency in the German language: A study of pupils in grade 7* 8
 - 1.1 Introduction 8
 - 1.1.1 Skills included in the tests 8
 - 1.1.2 The subject-matter content of the tests 9
 - 1.2 Problems and method of analysis 9
 - 1.3 Measuring instruments 10
 - 1.4 Experiment groups 11
 - 1.5 Results 11
2. *Measuring proficiency in the German language: A study of pupils in grade 9* 13
 - 2.1 Introduction 13
 - 2.2 Problems and method of analysis 13
 - 2.3 Test construction and test data 14
 - 2.4 Experiment groups 14
 - 2.5 Results 15
3. *Measuring language proficiency: A study testing hypotheses concerning a factor structure* 21
 - 3.1 Introduction 21
 - 3.2 The ACOVS programs 21
 - 3.3 Testing a hypothetic factor structure 22
 - 3.4 Results 22
4. *Measuring the effect of the UMT project's teaching system "Deutsch"* 24
 - 4.1 Introduction 24
 - 4.2 Problems 24
 - 4.3 Experiment groups 25
 - 4.4 Measuring instruments 26
 - 4.5 The reliability and validity of the sub-tests 26
 - 4.6 Results 27

References 30

Acknowledgements

These studies have been conducted at the Department of Educational and Psychological Research, School of Education in Malmö and have been made possible by funds from the National Board of Education.

I want to thank the head of the Institute Professor Åke Bjerstedt and acting professor Docent Olof Magne for their interest in my work.

In particular I want to thank Associate Professor Ebbe Lindell who in his stimulating way has highly motivated me to carry out these studies and I thank him for valuable discussions during the work.

I also want to thank all my colleagues in the UMT-project, Lektor Bertil Ekholm-Erb, Adjunkt Bertil Engh, Adjunkt Lennart Ericsson, Mrs. Inger Grönvall, Fil. lic. Per Hall, Fil. kand. Kerstin Naucér, Adjunkt Mats Olsson and Mrs. Britt Persson, without whose assistance these studies could not have been carried out. My sincere thanks also to those outside the project staff who have helped me with corrections and analyses of the measuring instruments, Docent Bernt Larsson for his advice in questions of research methodology and of statistical analysis, Fil. kand. Agneta Sternerup who made the data computing operations and Gillian Nilsson, B.A., for translating parts of this work into English.

Finally I also want to thank those pupils and their teachers who kindly took part in the extensive testing procedures.

Malmö, March, 1972

Horst Löfgren

Introduction

One of the goals of the research and development project UMT (Swedish: *Undervisningsmetodik i tyska*; English: Teaching methods in German) includes the scientific construction of a study material for the teaching of German to beginners. When the material was considered sufficiently developed for practical use, it went into its final phase, which was a comparison of the UMT material's effect on learning with that of other existing study materials. In such comparisons the actual measuring instruments are extremely important. The instrument used must measure the total language proficiency as comprehensively and impartially as possible. It should not be adapted for a particular study material or a particular teaching method, but should be applicable on a wider scale. What then is the structure of language proficiency and how is it to be measured as comprehensively as possible? A study of the literature on the subject showed that so far there was no definite answer to this question.

As a contribution to the understanding of the structure of language proficiency and of how it can be measured as comprehensively as possible, a study was conducted, starting from the goals of language teaching and a theory on the relationship between language variables and integrated skills. This study, which was carried out in grade 7 (Löfgren, 1969) in the teaching of German, was then followed up by a similar analysis in grade 9 (Löfgren, 1972 a; Löfgren, 1972 b). The results of the first investigation were then used as the basis for the construction of measuring instruments for studying the effects of the UMT project's study material (Löfgren, 1972 c).

1. Original report: Löfgren, H. Mätningar av språkfärdighet i tyska: En undersökning på elever i årskurs 7. /Measuring proficiency in the German language: A study of pupils in grade 7./ *Pedagogisk-psykologiska problem*, No. 84, Malmö: School of Education, 1969

1.1 Introduction

1.1.1 Skills included in the tests

Earlier analyses of language proficiency have often dealt with the problem of prediction. Attempts have been made to extract factors which are relevant to success in learning a second language (Carroll, 1958; Smith & Berger, 1968). Lado (1962) has discussed in detail the construction of suitable language tests. He has tried to break down language proficiency into smaller components in order to find common elements. Lado's idea is that language consists of elements which constitute variables which in their turn are not independent but are integrated into the skills: speak, listen, read, write and, possibly, translate.

Smith and Berger share Lado's opinion that the skills can be broken down into common variables. "All skills make use of the common core of structure, morphology and vocabulary" (Smith & Berger, 1968, p. 70).

Scherer and Wertheimer also take up the question of whether it is suitable to regard language proficiency as being composed of four different skills, or possibly six if translation in both directions is included. The division into listen, speak, read and write, they say, is strongly rooted in the world of the language teacher, as is the difference between active and passive ability. For decades the language teacher has spoken of language proficiency in just these terms. It is possible, however, that empirical studies would show that proficiency in a foreign language ought not to be divided into these a priori categories.

"Factor analytic approaches would be still more powerful in helping to decide what separable aspects can be empirically demonstrated to compose proficiency in a foreign language. Had computer capacity permitted, intercorrelations among all the subtests of all the proficiencies would have yielded direct evidence on the question of whether the traditional fourfold classification makes empirical sense. Perhaps in the future time and funds will become available for us to undertake such analyses." (Scherer & Wertheimer, 1964, p. 113.)

The curricula for the comprehensive and vocational training schools and the "gymnasium" in Sweden show clearly how total language proficiency is seen as

being made up of different skills, each of which should be tested separately to give as impartial a picture as possible of a student's achievements. They describe how student achievements should be assessed, but give no practical examples. The Swedish standard tests for the comprehensive school and the central tests for the "gymnasium" are aids in assessing what the students can do. The purpose of these tests is to produce norms for deciding marks and they are intended to provide a yard-stick for the entire country.

To summarize, we have lacked a synthesis of language proficiency based on empirical studies. Language ability has usually been divided into the skills: read, write, speak, hear and translate. If these skills are of different natures, as Brooks thinks, it ought to be possible to test them separately or in different combinations (Brooks, 1960). If each skill can be divided into component parts, such as the variables suggested by Lado, pronunciation, grammatical structure and vocabulary, there ought to be tests which measure these components.

In the national curricula, total language proficiency is seen as consisting of different skills, each of which should be tested separately so as to give as fair a picture as possible of a student's achievements. The standard tests and the central tests are naturally a consequence of the view of language proficiency taken in the curricula. The reason why oral production is not included in these tests is that this has been thought to be impossible for practical reasons.

1.1.2 The subject-matter content of the tests

We come now to the question of what language material the tests should contain. A distinction must again be made here between achievement and proficiency tests. If we want a test based directly on the course studied, then the objectives of this course will form the basis of the test construction. If we instead want a general language test independent of any course, then the material used can be decided by the objectives of the language teaching, as stated in, for example, curricula, text-book analyses, frequency studies and mistake analyses.

1.2 Problems and method of analysis

From the introductory discussion, the following problems emerge:

1. What different factors make up language proficiency?
2. Which measurements should be included in a test battery intended to measure total language proficiency as comprehensively as possible?
3. Are the different factors of which language proficiency consists influenced by teaching method?
4. Are the measurements included in a test battery, which is intended to measure total language proficiency, dependent on the teaching method?

To answer these questions we have used a factor-analytic method. We have made use of the standard program BMD03M (Dixon, 1967), which uses the principal component solution with varimax rotation.

1.3 Measuring instruments

The starting-point chosen for the test construction is the curriculum and a survey of skills and variables compiled on the basis of Lado's ideas (1962). Box 1.1 summarizes what the various sub-tests are intended to measure. A survey comparing the composition of the test battery with the curriculum is presented in Box 1.2.

Box 1.1 Survey of what the different sub-tests are intended to measure.

<i>test no.</i>	<i>test</i>	<i>intended to measure</i>
1	Listening 1	ability to understand spoken language
2	Listening 2	ability to understand spoken language
3	Reading comprehension 1	ability to understand written text
4	Reading comprehension 2	ability to understand written text
5	Write 1	ability to write the language
6	Write 2	ability to write the language
7	Speak 1	ability to speak the language
8	Speak 2	ability to speak the language
9	Translate 1	ability to translate from the target language to the mother tongue
10	Translate 2	ability to translate from the mother tongue to the target language
11	Vocabulary 1	the passive vocabulary
12	Vocabulary 2	the active vocabulary
13	Vocabulary 3	the active vocabulary
14	Pronunciation 1	phoneme pronunciation
15	Pronunciation 2	intonation and stress
16	Pronunciation 3	phoneme pronunciation, intonation and stress
17	Fluency 1	fluency in target language
18	Fluency 2	fluency in mother tongue; Thurstone's w-factor
19	Intelligence 1	the verbal factor; Thurstone's v-factor
20	Intelligence 2	power of reasoning; Thurstone's r-factor
21	Mark in German	total language proficiency as assessed by the school
22	Grammar test 3	grammatical structure; verbs
23	Grammar test 4	grammatical structure; verbs
24	Grammar test 9	grammatical structure; verbs
25	Grammar test 1, 2, 6, 7	grammatical test; case

Box 1.2 A comparative presentation of the objectives laid down in the curriculum for the comprehensive school and the content of the test battery.

<i>The wording of the curriculum</i>	<i>Content of the test battery</i>
The task of the teaching is to make the pupils . . .	
familiar with a limited, central stock of words,	vocabulary of various kinds
expressions and grammatical patterns and to found a good pronunciation, so that they can catch and understand clear speech,	elementary case and tense structures pronunciation tests listening to sentences and texts
read and understand easy text, express themselves orally in simple, everyday situations	silent reading of sentences and texts colloquial speech in various situations and with different demands
to some extent express themselves in writing	written production of various kinds

1.4 Experiment groups

The investigation was carried out in a number of classes in Malmö at the end of the seventh grade. Those participating were 7 classes (experiment group), which had been taught with stenciled copies of the UMT study material system and 8 classes (control group), which had been taught with other textbooks. All students received 5 hours' teaching in German a week. Complete data were obtained from 130 students in the experiment group and 151 students in the control group.

1.5 Results

The analyses have shown that language proficiency can be seen as a rather closely-knit unit and the correlations between the different language tests are high. It is possible, however, to sort out certain aspects of language proficiency. There appear to be four main factors which are significant for language proficiency. These have been named knowledge of words and structures, intelligence, pronunciation and fluency.

Thus the traditional division into speak, read, write and hear does not reflect the factorial composition of language proficiency. Proficiency in vocabulary and structures has proved to be a central factor. Hence, simple vocabulary and grammar tests can give a very accurate measurement of language proficiency. The factor structures were not quite the same in the two groups.

Pronunciation, intelligence and language fluency factors are more similar in the two groups.

The sub-tests showing the highest loadings on the first factor were the grammar tests and the two active vocabulary tests. High loadings were also registered for Write 1, Translate 2 and Speak 1, all of which test active ability.

The two intelligence tests together with Listening 2 and Reading comprehension 2 show the highest loadings in the second factor, which was interpreted as an intelligence factor.

The third factor, the pronunciation factor, only has high loadings in Pronunciation 1 and 2.

In the fourth factor, Speak 2 and Write 2 have the highest loadings. Both these tests aim at measuring free production. The construction of the test implies a certain risk that we have obtained what is known as an instrument factor here, but for the time being we are interpreting this factor as a fluency factor.

The factor structures are so similar for the two groups of students that the same test battery can be used to measure language proficiency. Thus it has not been found that the different teaching methods under study have produced such different factor structures that they will influence the composition of the test battery. This is a most satisfactory result and if, by generalization, this result could be applied to other teaching methods as well, it would mean that the same battery could be used in assessing different groups of students.

The suggested test battery which was drawn up in order to measure total language proficiency as comprehensively as possible contained the following sub-tests:

Factor 1. Grammar test on verb conjugations and case declension

Vocabulary test

Written test of the type Write 1

Factor 2. Reading comprehension and listening comprehension tests

Factor 3. Pronunciation test

Factor 4. Test of oral production of the type Speak 2

2. Original report: Löfgren, H. Mätningar av språkfärdighet i tyska: En undersökning på elever i årskurs 9. /Measuring proficiency in the German language: A study of pupils in grade 9./ *Pedagogisk-psykologiska problem*, No. 158. Malmö: School of Education, 1972

2.1 Introduction

The aim of this study is to follow up the analysis made in grade 7 (Löfgren, 1969) and to penetrate the subject more deeply in an attempt to find the answers to some of the problems which could not be solved with the help of the previous investigation such as measuring oral production and free writing.

The analysis of language proficiency in grade 7 provided some interesting information on the structure of language proficiency and suggested different tests which could be used for assessing the students' language proficiency as comprehensively as possible. It was considered essential to continue with similar analyses in other grades and possibly with other target languages. Thus this report is an analysis of the proficiency in the German language of students in grade 9. The experiences gained in the previous language proficiency study have been taken into account in this analysis.

It was mentioned in the report on the study in grade 7 that not many relevant studies have been carried out in this area. In more recent years, however, a large number of articles have been written on the measurement of language proficiency and on how to construct such tests. The relationship between the different elements of language proficiency has not been taken up, however. At the most a language proficiency analysis is suggested, together with a few important areas requiring research (cf. Valette, 1966).

A few relevant studies in this research area that were not included in the first report are discussed in this one. They include the results of an investigation carried out by Gardner and Lambert (1965). The results of their study in the learning of French are very similar to those of our earlier study in grade 7.

2.2 Problems and method of analysis

This study of language proficiency is intended to find answers to the following problems:

1. What is the factorial structure of language proficiency?

2. To what extent can measurements of language variables and the parts of the integrated skills that can be measured more objectively predict what the students can achieve in
 - a) free written production
 - b) free and specified oral production?
3. To what extent can measurements of language variables be used to predict the students' achievements in the integrated skills?
4. Which measurements should be included in a test battery which aims at measuring total language proficiency as comprehensively as possible?

For analysis of the above problems, the following methods have been chosen:

Problem 1. Factor analysis BMD03M (Dixon, 1967)

Factor analysis ACOVS (Jöreskog et al., 1970)

Problem 2. Multiple regression analysis BMD02R (Dixon, 1967)

Problem 3. Canonical analysis BMD06M (Dixon, 1967)

Problem 4. Conclusions from points 1—3 above

2.3 Test construction and test data

As in the study in grade 7, the starting point taken for the test constructions has been the curriculum and the survey of skills and variables presented in the previous study (Löfgren, 1969, p. 7). A summary is given below of what the different sub-tests in grade 9 are intended to measure.

Since the different language tests have been constructed on the basis of an analysis of the goals of language teaching and the relationship between language variables and integrated skills, the test battery can be said to have a high content validity.

The reliability of the tests must be considered satisfactory (.71—.92). Among the less reliable tests are those measuring listening comprehension. Earlier investigations have already shown that it is difficult to get reliable measurements from listening comprehension tests (Löfgren, 1969). Moreover, the sub-tests Vocabulary 1, Read 3, Intelligence 3, Translate 1 and Speak 3 appear to be less homogeneous. The sub-tests have been validated against the marks in German in grade 9. The validity coefficients for the language tests lie on the expected level (.50—.70).

2.4 Experiment groups

The study was carried out at the end of May 1970. From the population of grade 9 students in Malmö learning German 4 hours a week, a group was chosen consisting of 156 pupils from 3 different schools.

Box 2.1 Survey of what the different sub-tests are intended to measure.

<i>test no.</i>	<i>test</i>	<i>intended to measure</i>
1	Mark in German	total language proficiency
2	Vocabulary 1	passive vocabulary (translation of a single word in a German sentence)
3	Vocabulary 2	active vocabulary (filling in a missing word in a German sentence)
4	Vocabulary 3	active vocabulary (opposites)
5	Read 1	comprehension of single sentences
6	Read 2	comprehension of content of text extracts
7	Read 3	comprehension of written instructions
8	Write 1	ability to form sentences in German with given words
9	Write 2	ability to express oneself in writing in connection with given stimulus situations
10	Write 3	free writing in connection with a picture
11	Intelligence 1	verbal ability (v factor)
12	Intelligence 2	logical-inductive and verbal ability (i and v factor)
13	Intelligence 3	verbal ability (v factor)
14	Grammar 1	case declension (following preposition)
15	Grammar 2	verb forms
16	Translate 1 a	ability to translate from target language to mother tongue
17	Translate 1 b	ability to translate from target language to mother tongue
18	Translate 2 a	ability to translate from mother tongue to target language
19	Translate 2 b	ability to translate from mother tongue to target language
20	Pronunciation 1	intonation and stress
21	Pronunciation 2	pronunciation of phonemes
22	Listening 1	comprehension of oral instructions
23	Listening 2	comprehension of sentences
24	Listening 3	comprehension of content of stories
25	Speak 1	ability to converse
26	Speak 2	ability to express oneself in simple, everyday situations
27	Grammar 3	case declension (object)
28	Speak 3	free oral production
29	Pronunciation 3	pronunciation, intonation and stress

2.5 Results

Problem 1. What is the factorial structure of language proficiency?

As in the previous study in grade 7, the results show that language proficiency

is a closely-knit unit and the correlations between the language tests are often high. The pronunciation tests, however, show only average connections with the other language tests. The most central language tests are certain of the writing, translation, vocabulary and grammar tests. It is of particular interest to find, as with the study of grade 7, that the translation tests have such high correlations with the other language proficiency tests. So translation tests which have often been harshly criticized are nevertheless not too bad a gauge of the students' language proficiency.

To facilitate comparison of the results from the studies in grades 7 and 9, a re-analysis has been made of the data from the former study. In this re-analysis the test battery was better balanced as to the different parts of language proficiency.

The result of the re-analysis corresponds well with the original one. The intelligence test divided into two factors in the experiment group, but not in the control group. The most interesting point, however, is that we have not been able to find any language fluency factor here, the factor that in the original study we suspected could be an instrument factor. The first factor can be called a vocabulary and grammar factor. Here we have the highest loadings on the sub-tests assessing vocabulary and grammatical correctness, namely Vocabulary 2 + 3, Write 1, the grammar tests and the two oral tests. The other factor, verbal intelligence, has in addition to Intelligence 1 the highest loadings on the sub-tests measuring passive language proficiency. In the third factor are to be found the test of reasoning ability and the grammar tests. The fourth factor is not easily interpretable, but the two pronunciation tests show the highest loadings.

As a result of the studies carried out in grade 7, we now have some idea of the structure of language proficiency, and we have reasonably clear hypotheses for further studies both here and on other educational levels. We suppose that the language proficiency can be described in the following way:

Factor A. Vocabulary and grammar factor

Factor B. Intelligence factor

Factor C. Pronunciation factor

Our next aim is to test the hypothesis on the above factor structure in the present study.

Since the ACOVS program was still not available, a factor analysis was carried out on the material from grade 9 according to BMD03M (Dixon, 1967).

In this factor analysis three or possibly four interpretable factors were extracted. The first factor (A) traps a large amount of the common variance. The only language tests that do not have high loadings on this factor are the pronunciation, reading comprehension and listening comprehension tests. Since

the vocabulary and grammar tests show very high loadings, this factor A can with good reason be regarded as a grammar and vocabulary factor. In assessing the tests intended to measure ability to write and translate, it was in particular the choice of words and grammatical correctness that were judged.

Factor B is quite clearly an intelligence factor and here we find the reference tests that measure verbal ability. As in the case of the analysis in grade 7, we find that the reading comprehension test links up with the three intelligence tests. The three listening comprehension tests, on the other hand, do not show any substantial loadings on this factor.

The three pronunciation tests differ from the other parts of language proficiency and this again is in complete agreement with the hypothesis. We interpret factor C quite simply as a pronunciation factor.

The fourth factor (D) is more difficult to interpret. It is possible that what we have picked up here is a so-called method factor, which has emerged because of a particular test situation. The factor can also, however, reflect a genuine part of language proficiency. It can be a listening comprehension factor or, more probably, a passive language proficiency factor.

To summarize, the hypothesis that was drawn up concerning the factor structure in grade 9 seems rather probable. It is doubtful, however, whether intelligence and passive language proficiency should be linked, or whether we here have two more widely separated areas. As was mentioned earlier, our intention is to test the hypothesis statistically with the help of special programs as soon as these are available. This analysis is discussed later.

Problem 2. To what extent can one use measurements of language variables and the parts of the integrated skills that are more easily measured objectively to predict the students' achievements in

- a) free written production
- b) free and specified oral production?

In this and the following section, we study the relationship between different parts of language proficiency, in order to draw particular attention to the possibility of using simple, more objective tests to predict the parts of language proficiency that are more difficult to measure. This first section will be devoted to studying how the students' free oral and written production can be predicted. The other two separate tests for oral production are also included here since it is in particular the prediction of oral fluency that is of great practical interest.

In the section that follows, the possibility of predicting the four integrated skills, hearing, speaking, reading and writing, will be studied. There the result will be predicted for the skill as a whole, as it has been defined by the sub-tests included. The third language variable, "pronunciation", has not been included

Box 2.2 The prediction of Speak 1—3 and Write 3 on the basis of a selection of the more objective sub-tests.

Independent variables	Dependent variables	Multiple corr.
a) Vocabulary 3, Grammar 1, Write 1, Listening 1	Speak 1	.78
b) Vocabulary 3, Grammar 1, Write 1, Listening 1	Speak 2	.80
c) Vocabulary 3, Grammar 1, Write 1, Listening 1	Speak 3	.70
d) Vocabulary 3, Grammar 1, Write 1, Listening 1	Write 3	.81

in these two sections. The reason for this is that it has proved to form a separate factor. This means that problems 2 and 3 have partly been simplified. As a result of this simplification, however, prediction of oral fluency means prediction of the students' ability to express themselves in the foreign language, without any importance being placed on the pronunciation of what is said.

The method used here to ascertain the efficiency of certain language tests in predicting free written production and free and specified oral production has been step-wise multiple regression analysis (Dixon, 1967).

The analyses showed that the achievements of the students in free and specified oral production and free written production can be satisfactorily predicted. The box above shows a summary survey of the results.

It should perhaps be added that the above results as well as the other results of the factor analysis are consequences of the current way of teaching foreign languages. You cannot, for instance, train only vocabulary and grammar and then expect a good performance in oral skills.

Problem 3. To what extent can one use measurements of language variables to predict the students' achievements in the integrated skills?

The factor-analytical studies of the structure of language proficiency have shown that the factor that we have called the vocabulary and grammar factor is very central. Most of the language tests show loadings on this factor. In the same way we have seen in the previous section the great importance of the vocabulary and grammar tests. The starting point for the entire study of language proficiency has been an analysis of the relationship between language variables and the integrated skills. Now that these language variables have proved to be of such importance, it is of interest to investigate to what extent one can use these alone to predict integrated skills. In other words, we shall try to measure the connection between some vocabulary and grammar tests and some tests on the integrated skills, reading, writing, listening and speaking.

Box 2.3 · Prediction of the integrated skills on the basis of some vocabulary and grammar tests.

First set of variables (u)	Second set of variables (v)	r_{uv}
a) Vocabulary 1—3 Grammar tests 1—3	Read 1—2	.84
b) Vocabulary 1—3 Grammar tests 1—3	Write 1—2	.92
c) Vocabulary 1—3 Grammar tests 1—3	Listening 1—3	.78
d) Vocabulary 1—3 Grammar tests 1—3	Speak 1—2	.85

The method chosen for shedding light on the problem of prediction is canonical correlation analysis. The method is available in a standard program (Dixon, 1967). In the analyses, one group of variables has been the three vocabulary and the three grammar tests while the other group has consisted of tests of the integrated skills.

The box above gives a summary of the results. They establish that it is possible to assess the language proficiency of the students simply and impartially. By giving a few vocabulary and grammar tests, we can with a fair degree of reliability grade the students in relation to one another in the integrated skills.

Problem 4. What measurements should be included in a test battery intended to measure total language proficiency as comprehensively as possible?

The combined results of the three previous analyses can be used in answering this fourth problem. In order to make the results directly applicable for language teachers working in the schools, the sub-tests will be discussed as they stand, without any adjustments via calculating operations.

If we compare the summary of the factor analysis, the summary of the regression analysis and in addition the canonical correlation analyses, we find that the test battery below can very well be used and expected to produce reliable and valid results. This is not to say that these sub-tests cannot be improved upon as long as they still measure the same things. Moreover, when compiling the test battery, one should keep in mind its validity from the students' point of view. They should feel that the sub-tests are relevant to the aims of the language teaching.

Box 2.4 Suggestion for a test battery which is to measure total language proficiency as comprehensively as possible.

<i>Sub-test</i>	<i>Comment</i>
Vocabulary 3	The test measures active vocabulary. The test should preferably be a straightforward check on words without additional distracting elements such as words placed in sentences. If one wishes to make it uni-lingual, opposites can be used.
Grammar tests 1, 2	The grammar tests can deal with case declension and conjugation of verbs. The results can possibly be summarized.
Write 1	The test can be made rather easy to mark and this test has shown high correlations with the other parts of language proficiency.
Read 2	Reading comprehension has proved to be a rather specific skill correlated with intelligence. It is possible that Listening 3 also measures this element.
Listening 1 (3)	If separate listening comprehension tests are desired, Listening 1 is suitable for inclusion because of its relation to oral proficiency. Listening 3 rather tests comprehension as a whole, and is constructed analogously with Read 2. If no listening comprehension test is included, one should instead have a test of passive vocabulary, such as Vocabulary 1.
Pronunciation 1, 2	Pronunciation has proved to be a rather specific language skill and should therefore be tested separately.

3. Original report: Löfgren, H. Mätningar av språkfärdighet: En hypotesprövande undersökning av en faktorstruktur. /Measuring language proficiency: A study testing hypotheses concerning a factor structure./ *Pedagogisk-psykologiska problem*, No. 161. Malmö: School of Education, 1972

3.1 Introduction

Starting from the goals of language teaching and an analysis of the relationship between language variables and integrated skills, studies have been made of the factorial structure of proficiency in the German language in grades 7 and 9 and of how this can best be measured.

On the basis of the results of the analysis in grade 7 and the later re-analysis, a hypothesis was put forward about a general factorial structure of language proficiency. It was assumed that language proficiency could be seen as being built up of the following three factors:

Factor A. Vocabulary and grammar factor

Factor B. Intelligence factor

Factor C. Pronunciation factor

When the study in grade 9 was carried out (Löfgren, 1972), there were no methods available for testing the structural similarities. Since methods have now been worked out for making such comparisons and programs for these have been implemented on computers in Sweden, the division into factors given above can be tested statistically. This report is a continuation of the analysis in grade 9.

3.2 The ACOVS programs

The ACOVS programs (analysis of *covariance* structures) can among other things be used for different types of hypothesis-testing in factor-analytical studies. The general method for comparisons of covariance matrices may be found in an article in *Biometrika* (Jöreskog, 1970) and the ACOVS program used in this study is described in a manual (Jöreskog et al., 1970).

Since it was in this study that the program was put first into practical use in Sweden, the ACOVS program and the actual method involved is described in some detail in the original report.

3.3 Testing a hypothetic factor structure

It was assumed that the factor structure discovered for language proficiency should also apply to a new random sample of students. It should make no difference if the test variables were different, if only they were constructed for measuring the same area.

$$H_0: A = A_0$$

$$p \text{ (variables)} = 20$$

$$r \text{ (factors)} = 3$$

$$\text{level of significance } 1 \%$$

The factor loading matrix (A) is specified completely with limited values (constrained parameters), while the unique variance parts (ψ^2) are wholly free.

3.4 Results

The initial matrix is in our case identical with the correlation matrix. The factor loading matrix is specified to three factors and to three different values for the factor loadings. The unique variance parts are estimated with the help of the program.

The hypothesis is tested by comparing the initial matrix to the reproduced matrix (goodness-of-fit test).

$$\chi^2 = 341.10$$

$$df = \frac{p(p+1)}{2} - m = 187$$

$$\text{where } p = \text{no. of variables}$$

$$m = \text{no. of estimated parameters}$$

$$P(\chi^2 \geq 341.10 | H_0) < .001$$

The hypothesis must be rejected. With the specifications made, our model does not adequately describe the results gained empirically. The reason for this can be either that three factors are not enough to describe language proficiency completely or that the specifications made concerning the factor loadings are not correct.

A control run with all the parameters free also produced a significant χ^2 -value, which shows that the number of factors is insufficient. This result does not mean that the three hypothetic factors are necessarily incorrect, but that we have shown that these three factors do not suffice to describe the whole truth about the structure of language proficiency.

In order that we might gain further information about how well language proficiency can be described with the help of four factors, a new analysis was carried out. The difference between the reproduced correlation matrix based on these values and the initial matrix was then considerably reduced and we

could not on the chosen level of significance reject the null hypothesis ($P|H_0 > .01$).

To summarize, we have not been able in this study to prove the assumption that language proficiency can be broken down into three main factors. The three factors, called vocabulary and grammar factor, intelligence factor and pronunciation factor, have in this study proved to be inadequate to give a complete description of language proficiency. There is reason to believe, however, that the three factors describe a very substantial part of the truth and the addition of a fourth factor in a follow-up study confirmed this belief.

4. Original report: Löfgren, H. Effektmätning av UMT-projektets undervisningssystem "Deutsch". /Measuring the effect of the UMT project's teaching system "Deutsch"./ *Pedagogisk-psykologiska problem*, No. 159. Malmö: School of Education, 1972

4.1 Introduction

The main problem in this study was to see what result teaching with the UMT project's study material had produced (*Deutsch*, 1972). We wished to compare our teaching system with other study materials used in the same grade. In addition, the data collected from all study materials were used for a description of how much students know after studying German for one year in different courses. The results can later be used as norms for achievement for further studies in grade 7. Furthermore, when carrying out this investigation, we shed some light on the problem of how the teaching might be organized and on how great the influence of the teacher is contra that of the study material.

The report starts with a summary of the work carried out within the UMT project and a survey of the literature on international comparisons between different methods and materials systems in foreign language teaching.

Previous to the investigation reported here, a smaller study of the same kind was carried out during the school year 1967/1968 in connection with the study in grade 7. Our experiment group then worked with a stenciled version of the UMT project's study material, while the control group worked with some of the printed text-books that were being used in Malmö at that time. The number of students with UMT was 130, those with other study material 151. The test battery was compiled on the basis of a language proficiency analysis and an attempt to concretize the goals of language teaching as stated in the curriculum of 1962 (Löfgren, 1969).

In this comparison the primitive stenciled experimental version of the UMT study material produced such good results that continued revision of the material was considered well-motivated. With a better layout, even better results were to be expected.

For the aims and a description of the UMT teaching system, see the rather detailed review in the original report. See also Lindell (1969, 1970, 1971).

4.2 Problems

Problem 1 a. What effects on learning does the UMT material have in relation to other existing study materials taken as a whole?

Problem 1 b. What effects on learning can be achieved with the UMT study material in relation to the different existing study materials taken individually?

Problem 2. What level in the various language skills do the students in grade 7 achieve in the different courses (larger course 5 hours, smaller course 3 hours, combined course 4 hours)?

Problem 3. What effects on learning are achieved by the UMT material with different forms of organization?

Problem 4. What are the comparative roles of the teacher and the study material in deciding how much the students learn?

4.3 Experiment groups

The study was conducted in Malmö. It may be assumed that, as far as the problems stated above are concerned, Malmö is a relatively representative sample of the student and teacher populations.

The year the study was carried out, four different study materials were used in Malmö for the larger course in German. In the report these are called UMT, A, B, and C. In the smaller course, five different study materials were used, which are called UMT, D, E, F, and G. In the combined course, the UMT material was used.

It is debatable whether this study should not instead be seen as a total study in Malmö, in which case we should refrain from making any attempts to generalize to the populations above. If this investigation is seen as a total study, the significance testings that come later in the report naturally become unnecessary.

Students: 2012 of the approx. 2200 students taking German in grade 7.

Larger course (5 hours): 1168 students.

Smaller course (3 hours): 613 students.

Combined course (4 hours): 231 students.

Study material: All the study materials being used in Malmö during the school year 1968/69.

Larger course:

UMT: 6 schools, 462 students

A: 1 school, 87 students

B: 4 schools, 303 students

C: 4 schools, 315 students

Smaller course:

UMT: 2 schools, 83 students

D: 1 school, 30 students

E: 1 school, 51 students

F: 7 schools, 321 students

G: 3 schools, 128 students

Teachers: All 79 teachers teaching German in grades 7—9 of the comprehensive school in Malmö.

4.4 Measuring instruments

Data were collected on two main occasions. During the autumn term a series of preliminary testings were carried out. At the end of the following summer term, a series of post-testings were carried out, consisting of language proficiency tests and a test of attitude to German as a school subject. The following data have been collected:

Student data: Sex, year of birth, school, class, course, teacher.

Prel. testings: Intelligence: general IMU:t (Werdelin, 1958), verbal DBA:l (Härnqvist, 1960), inductive DBA:3 (Härnqvist, 1960); Attitudes: IMU:a (Werdelin & Ljung, 1968); Mark in grade 6 in Swedish and English.

Language tests: Reading comprehension, Vocabulary 1 and 2, Write 1 and 2, Speak, Listening comprehension.

Attitude measurements: Comparisons between pairs of different subjects.

A correlation analysis showed that only a few of the preliminary tests were of any interest. We were able to establish that the mark in English in grade 6 consistently showed a higher correlation with the measurement variables than the other two relevant preliminary tests, IMU: Instructions and DBA: Similarities.

The decision on which language tests should be included in the test battery was based partly on a study of the structure of language proficiency (Löfgren, 1969), and partly on the suggestions made at a conference with a reference group held in Malmö a month before the measurements of effect were carried out. One of the problems put in the language proficiency analysis was: "Which measurements should be included in a test battery intended to measure total language proficiency?" As a result of this investigation, a test battery was proposed that was based on a factor-analytic solution of language proficiency. Note that the tests are not identical with those mentioned earlier.

4.5 The reliability and validity of the sub-tests

In order to gain some idea of the reliability of the various sub-tests, we carried out a reliability assessment with the help of Cronbach's alpha coefficient (Rozeboom, 1966, p. 409—412).

The different sub-tests are obviously accurate in their measurements (.75—.86). Since we used the tests for group comparisons, the degree of accuracy is completely satisfactory. No validity data, for marks in the subject in the grade concerned, for example, were collected. The correlation with marks for similar language tests has proved to be around $r = .50—.70$ (see p. 14). In this case the way in which the different sub-tests have been constructed guarantees content validity. As was mentioned earlier, the tests were based on

an earlier analysis of language proficiency and on the decisions reached at a conference with a reference group.

4.6 Results

Problem 1 a. What effects on learning does the UMT material have in relation to other existing study materials taken as a whole?

Problem 1 b. What effects on learning can be achieved with the UMT study material in relation to the different existing study materials taken individually?

The problem was split in this way because we were primarily interested in comparing the UMT material with some kind of standard level. Putting all the other study materials together in one group was a suitable way of achieving this aim. The result in this group then shows the average level of achievement of the students. In addition this method of combining all the other materials leads to a greater number of classes being involved, which can reduce some possible sources of error, such as teacher influence. The disadvantage of combining them in this way is that possible differences between the groups of study materials may get evened out. For this reason a comparison was also made between the UMT material and each other individual study material.

In order to find an answer to the problem set, a series of single-factor experiments were carried out, with two or four levels in the factor. In comparisons of the measurement variables a significance level of 1 % was chosen. Since we were unable to distribute either study materials or teachers to the students at random, we considered it best to try as far as possible to investigate to what extent the groups and their teachers could be regarded as being on equal terms at the start. Since intelligence and school marks have proved to be essential factors, an investigation of these variables should tell us quite a lot about the comparability of the groups.

These analyses showed that the different groups of study materials could be considered as being equal in the initial stage.

In the main comparison in the larger course of the effects on learning achieved with the UMT material in relation to other existing study materials, the UMT material proved to function extremely well. In all the post-tests, seven language tests and an attitude test, the UMT group showed the best results compared to the other groups of study materials combined. In one of the vocabulary tests and the attitude test, however, the difference was not significant at the 1 % level. There were also considerable differences when the different study materials, UMT, A, B, and C, were compared individually. The estimation of ω^2 values (Hays, 1969, p. 381—384) also showed that the differences are educationally important, particularly for the sub-tests Vocab-

ulary 2, Listening 1 and Write 2. From the average values for the different groups of study materials, without taking statistical significance into consideration, we get the following results:

1. Students using the UMT project's study material achieved better results than the students using study material A in all five language tests (students using material A did not take part in the oral tests) and had in addition a more positive attitude to German as a subject.
2. Students using the UMT project's study material achieved better results than the students using study material B in six of the seven language tests and had in addition a more positive attitude to German as a subject.
3. Students using the UMT project's study material achieved better results than the students using study material C in six of the seven language tests and had in addition a more positive attitude to German as a subject.

Despite the fact that the UMT project's study material was not constructed for the smaller course, it proved to give positive results there as well. The UMT material has no difficulty in holding its own when compared to other study materials. Since the groups compared were relatively small, however, a certain amount of caution is called for with regard to our results, since there can have been other reasons than the study material for the variation. We have not, for example, been able to check the teacher influence.

Problem 2. What level in the various language skills do the students in grade 7 achieve in the different courses (larger course 5 hours, smaller course 3 hours, combined course 4 hours)?

The study of this problem serves two main purposes. Firstly, the results in the different courses can function as norms, i.e. the results describe how far one usually gets today in one year's German. The test results can also be used as a norm for later testings and assessments of study material. Secondly, this problem analysis can provide information for decision-makers when discussing the possibility of merging the larger and smaller courses.

The achievements of the groups differed greatly in the post-tests. The differences were consistently of a high significance and the ω^2 values were also very high, with the exception of the attitude measurements. Since the groups differed in the initial stage and had in addition been taught different amounts, the result was foreseen.

Even though there are big differences in the average values, there is also a certain, not unimportant, overlap in the spread of achievement. In other words, there are students with only 3 hours a week who have nevertheless done better than many of those taking 5 hours a week. This might be seen as evi-

dence of a wrong decision on the part of some students in the larger course. The students with extremely good results in the smaller course have perhaps chosen more from interest. This can naturally also be seen as a wrong decision. The combined course included in this study has only had 4 hours a week and seems in addition to have had a somewhat lower initial standard.

Problem 3. What effects on learning are achieved by the UMT material with different forms of organization?

The UMT project's study material was also tested in classes organized in somewhat different ways. Thus there was one group working with two teachers and an assistant teacher to every three classes and one group working with one teacher and one assistant teacher to every two classes. These two methods of organization were compared with the normal one of having one teacher per class. As a result of the comparisons it seems fairly safe to say that the changes in the circumstances under which the students were taught did not cause any deterioration in their achievements in a series of language tests. On the contrary the results seem so good that there is reason for continuing with similar experiments. Again, however, the groups compared were small so that the teacher influence can have introduced a considerable source of error.

Problem 4. What are the comparative roles of the teacher and the study material in deciding how much the students learn?

Different analyses have shown here that the level of achievement after one year's teaching varies enormously. These variations between classes are not always in line with the initial tendencies. There are teachers who despite a rather weak initial position make great progress with their classes and others whose classes despite a rather favorable initial position, do comparatively badly in the language tests. These substantial variations in the achievements of the different classes point to the teacher as being a very important factor.

In summarizing, this study of effect enables us to establish that the study material constructed by the UMT project can produce excellent results. Vigorous research and successive testings and try-outs of study materials obviously pay. But this study has also shown that there is still a great deal left to be done to modify and balance the role of the teacher and thereby lessen the variation between classes.

References

- Brooks, N. *Language and language learning*. New York: Harcourt, Brace & World, 1960.
- Carroll, J. B. A factor analysis of two foreign language aptitude batteries. *J. gen. Psychol.*, 1958, 59, 3—19.
- Deutsch*. Lund: Gleerups, 1972.
- Dixon, W. J. (Ed.) *Biomedical computer programs*. Berkeley, Los Angeles: University of California Press, 1967.
- Gardner, R. C. & Lambert, W. E. Language aptitude, intelligence and second-language achievement. *J. educ. Psychol.*, 1965, 56, 191—199.
- Hays, N. L. *Statistics*. New York: Holt, 1969.
- Härnqvist, K. *Individuella differenser och skoldifferentiering*. Stockholm: SOU, 1960.
- Jöreskog, K. G. A general method for analysis of covariance structures. *Biometrika*, 1970, 57, 237—251.
- Jöreskog, K. G., Gruvæus, G. T. & van Thillo, M. ACOVS. A general computer program for analysis of covariance structures. *Research Bulletin*, Princeton: ETS, stencil, 1970.
- Lado, R. *Language testing*. Bristol: Longmans, 1962.
- Lindell, E. A system of study materials for the teaching of foreign languages: Principles and examples. *Didakometry & Sociometry*, 1969, 1, 91—124.
- Lindell, E. (Red.) Att arbeta med ett undervisningssystem: En presentation av UMT-projektets läromedel. /Working with a teaching system. A presentation of the UMT-project's study materials. / *Pedagogisk-psykologiska problem*, Nr 106, 1970.
- Lindell, E. *Främmande språk i grundskolan*. Lund: Gleerups, 1971.
- Löfgren, H. Mätningar av språkfärdighet i tyska: En undersökning på elever i årskurs 7. /Measuring proficiency in the German language: A study of pupils in grade 7. / *Pedagogisk-psykologiska problem*, Nr 84, 1969.
- Löfgren, H. Mätningar av språkfärdighet i tyska: En undersökning på elever i årskurs 9. /Measuring proficiency in the German language: A study of pupils in grade 9. / *Pedagogisk-psykologiska problem*, Nr 158, 1972 a.
- Löfgren, H. Mätningar av språkfärdighet: En hypotesprövande undersökning av en faktorstruktur. /Measuring language proficiency: A study testing hypotheses concerning a factor structure. / *Pedagogisk-psykologiska problem*, Nr 161, 1972 b.
- Löfgren, H. Effektmätning av UMT-projektets undervisningssystem "Deutsch". /Measuring the effect of the UMT-project's teaching system "Deutsch". / *Pedagogisk-psykologiska problem*, Nr 159, 1972 c.
- Rozeboom, W. W. *Foundations of the theory of prediction*. Homewood, Ill.: The Dorsey Press, 1966.

- Scherer, G. & Wertheimer, M. *A psycholinguistic experiment in foreign-language teaching*. New York: McGraw-Hill, 1964.
- Smith, P. D. & Berger, E. *An assessment of three foreign language teaching strategies utilizing three language laboratory systems*. Washington, D.C.: U.S. Department of health, education, and welfare, 1968.
- UMT: *Tyska 7*. Stockholm: SÖ-förlaget, 1968—69.
- Valette, R. M. *Modern language testing*. New York: Harcourt, Brace & World, 1964.
- Werdelin, I. *The mathematical ability*. Lund: Håkan Ohlssons boktryckeri, 1958.
- Werdelin, I. & Ljung, B. O. *IMU:a*. Malmö: Lärarhögskolan, 1968.

Published by the University of Lund, the Malmö School of Education, and the Society of Psychology and Pedagogy:

STUDIA PSYCHOLOGICA ET PAEDAGOGICA

SERIES ALTERA

1. Edlund, Sven: Diskussionen om begåvningsurvalet under reformations- och stormaktstiden. I. 1947.
2. Agrell, Jan: Nyinrättade professurer inom filosofiska fakulteten i Lund under 1800-talets första hälft. 1949.
3. Smith, Gudmund: Psychological Studies in Twin Differences. 1949.
4. Husén, Torsten: Om innebörden av psykologiska mätningar. Några bidrag till psykometrikens metodlära. 1949.
5. Smith, Gudmund: Interpretations of Behaviour Sequences. 1952.
6. Johannesson, Ingvar: Studier av sociala relationer mellan barn i folkskoleklasser. 1954.
7. Kragh, Ulf: The Actual-Genetic Model of Perception-Personality. 1955.
8. Bjerstedt, Åke: Interpretations of Sociometric Choice Status. 1956.
9. Werdelin, Ingvar: The Mathematical Ability. Experimental and Factorial Studies. 1958.
10. Werdelin, Ingvar: Geometrical Ability and the Space Factors in Boys and Girls. 1961.
11. Borg, Gunnar A. V.: Physical Performance and Perceived Exertion. 1962.
12. Lindell, Ebbe: Den svenska rättskrivningsmetodiken. 1964.
13. Bokander, Ingvar: Stereoscopically Induced Perceptual Conflict between Facial Photographs. A Discussion and Summary of Results. 1968.
14. Bjerstedt, Åke: Twelve Years of Educational and Psychological Research in Sweden: A Bibliography. 1968.
15. Larsson, Bernt: Bayes Strategies and Human Information Seeking. 1968.
16. Bjerstedt, Åke: Educational Technology. Instructional Programming and Didakometry. 1972.
17. Löfgren, Horst: The Measurement of Language Proficiency. 1972.

CWK GLEERUP
LUND / SWEDEN